Desarrollo de un Repositorio para la Gestión de Datos Primarios Georreferenciados

Gustavo Samec[1,3], Fernando Pap[1,3], Nicolás Gallia[1,3], Juan Emilio Sala[2], Flavio Quintana[2], Renato Mazzanti[1,3]

[1] Unidad de Gestión de la Información (UGI), CCT CENPAT-CONICET, Blvr. Brown 2915, U9120ACD, Puerto Madryn, Argentina
{gsamec, pap, ngallia, renato}@cenpat-conicet.gob.ar
[2] Instituto de Biología de Organismos Marinos (IBIOMAR), CCT CENPAT-CONICET Blvr. Brown 2915, U9120ACD, Puerto Madryn, Argentina
{juansala, quintana}@cenpat-conicet.gob.ar
[3] Laboratorio de Investigación en Informática (LINVI), Universidad Nacional de la Patagonia San Juan Bosco, Facultad de Ingeniería, Puerto Madryn, Argentina Blvr. Brown 3051, U9120, Puerto Madryn, Argentina

Abstract. El presente trabajo describe el desarrollo de un sistema que provee una solución al manejo de grandes volúmenes de datos primarios generados por dispositivos adquisidores de datos (dataloggers) y posibilita su búsqueda en un Repositorio Institucional por medio de georreferencias. En particular para esta implementación se utilizaron datos de pingüinos y cormoranes donde se registran sus datos georreferenciados y archivos multimedia asociados. La solución propuesta permite la catalogación de los registros, incorporación de metadatos, almacenamiento, control de acceso y consultas. El sistema integra DSpace, PostGIS y Leaflet.

Keywords: Datalogger, DSpace, Leaflet, PostGIS.

1 Introducción

La utilización de dataloggers aplicada a la investigación y en particular al comportamiento y ambiente donde habitan determinadas especies de animales permite obtener datos georreferenciados de calidad [1] valiosos para su estudio [2]. El avance tecnológico ha permitido un grado de miniaturización, calidad de los sensores y reducción de costos, impensados tiempo atrás, lo que ha posibilitado una amplia utilización de los mismos [3]. Ahora bien, surge la necesidad de gestionar el gran volumen de datos que generan para preservarlos y facilitar su búsqueda [4].

En particular, se necesita gestionar un volumen importante de datos registrados por dataloggers colocados en Pingüinos de Magallanes (Spheniscus magellanicus) y Cormoranes Imperiales (Phalacrocorax atriceps) por el grupo de investigación del Laboratorio de Ecología de Predadores Tope Marinos (LEPTOMAR) perteneciente al Instituto de Biología de Organismos Marinos (IBIOMAR) del CONICET, haciéndolos

accesibles a todos sus integrantes y facilitando su búsqueda por distintos criterios para recuperar el material registrado.

Los dataloggers utilizados se adosan a los cuerpos de determinados individuos, poseen una memoria interna no volátil y luego de un tiempo de funcionamiento se recuperan para descargar los datos generados. Los modelos utilizados actualmente generan archivos de datos georreferenciados en distintos formatos, secuencias de fotos y videos. Cada datalogger registra alrededor de 5 GB (gigabyte) de datos y cerca de 100 GB de material multimedia por muestreo. Cada año se realizan alrededor de 60 muestreos. Aproximadamente unos 6 TB (terabyte) de datos anuales.

Todos los datos generados por los dataloggers junto con bitácoras y documentación relacionada al proceso de instalación y recuperación de los dispositivos, se encuentran actualmente dispersos en distintos medios de almacenamiento y no cuentan con un sistema estándar para clasificarlos y facilitar su búsqueda. Se destaca la existencia inicial de un gran volumen de datos adquiridos a lo largo de los últimos 15 años y el crecimiento periódico de los mismos, lo que hace necesario contar con herramientas para su gestión.

El presente trabajo describe los desarrollos realizados integrando diferentes herramientas, que aportan una solución a este problema.

Luego de analizar varias alternativas para centralizar los datos y dejarlos accesible se optó por utilizar nuestro Repositorio Institucional (RI) implementado en DSpace [5], el cual permite gestionar adecuadamente el almacenamiento de los mismos.

Si bien DSpace cuenta con herramientas potentes de búsqueda, se encontró que éstas no satisfacen las necesidades requeridas por el grupo de investigación, por ejemplo extraer todos los registros pertenecientes a una determinada área geográfica. Para ello fue necesario extender el funcionamiento de DSpace y desarrollar un sistema integrado al mismo, que permite hacer consultas avanzadas y facilita la recuperación de los datos por su georreferencia.

En este trabajo exponemos las diferentes etapas de diseño e implementación desarrolladas. Las secciones 2.1, 2.2 y 2.3 presentan el diseño y la arquitectura propuesta y en las secciones 2.4, 2.5 y 2.6 se detallan los aspectos más relevantes del desarrollo en el repositorio, base de datos y mapa interactivo. Finalmente en la sección 3 se incluyen los resultados obtenidos y trabajos futuros.

2 Desarrollo

2.1 Análisis de Requerimientos

En principio se encontraron tres requerimientos que el sistema debe proveer:

(a) Almacenar datos georreferenciados en el repositorio: en el proceso de carga, cada unidad de información (ítem) debe contener todos los archivos generados por el datalogger y toda la documentación relacionada con el mismo. Además se deben registrar en los metadatos del ítem, datos relacionados con el datalogger utilizado, datos biológicos del individuo

muestreado y datos que permitan realizar un postproceso de los archivos generado por el datalogger que contiene datos georreferenciados.

(b) Realizar búsquedas basadas en áreas geográficas: mediante una interfaz gráfica que muestra un mapa, trazar un rectángulo (o definirlo ingresando sus coordenadas). Luego el sistema buscará y presentará todas las derrotas (trayectoria de los individuos) que intersectan con el área del rectángulo.

(c) Realizar búsquedas basadas en metadatos: facilitar la búsqueda de ítems basado en uno a más de sus metadatos y combinaciones de los mismos.

2.2 Arquitectura

Este desarrollo tiene una arquitectura compuesta por capas (Figura 1). En la primera capa se encuentra el repositorio (DSpace) junto con los módulos desarrollados para la carga y procesamiento de archivos con georreferencias. En la segunda capa, la base de datos (PostgreSQL) mantiene la estructura y datos del repositorio junto con la extensión PostGIS para dar soporte a datos geográficos. La última capa está compuesta por la aplicación con interfaz gráfica, desarrollada en php y jQuery, que hace uso de las bibliotecas de JavaScript Leaflet y Mapbox.

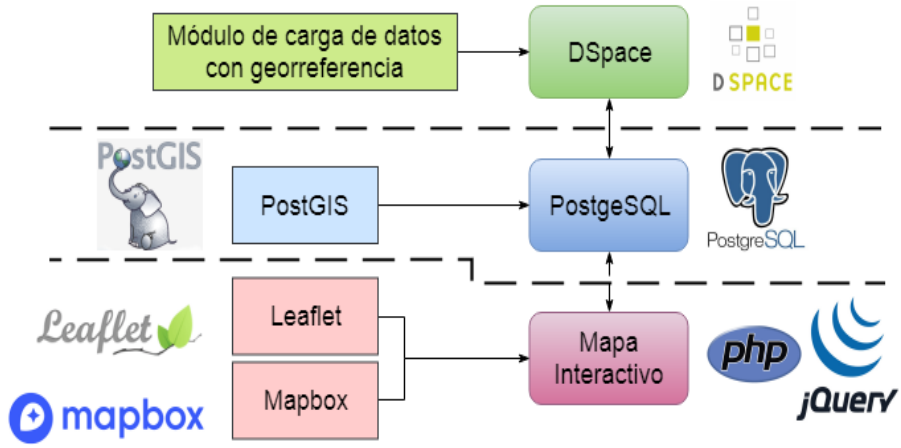

Figura 1. Arquitectura y tecnologías utilizadas.

2.3 Caso de Uso

El caso de uso en la figura 2 muestra las interacciones de los usuarios con el sistema y las capas de la arquitectura.

El usuario representado a la derecha es el encargado de cargar los datos primarios directamente al repositorio. Esta acción dispara subprocesos los cuales analizan y validan los datos, en particular las georreferencias, para ser cargadas en la base de datos.

El usuario representado a la izquierda realiza consulta de datos, el mismo interactúa directamente con la capa de presentación la cual posee distintas opciones para

seleccionar los datos a buscar. Cada selección genera una consulta específica en la base de datos y retorna los resultados encontrados en el repositorio, para luego visualizarlos de forma gráfica. También se visualizan los links que corresponden a los datos primarios para acceder, en forma directa, a los mismos en el repositorio.

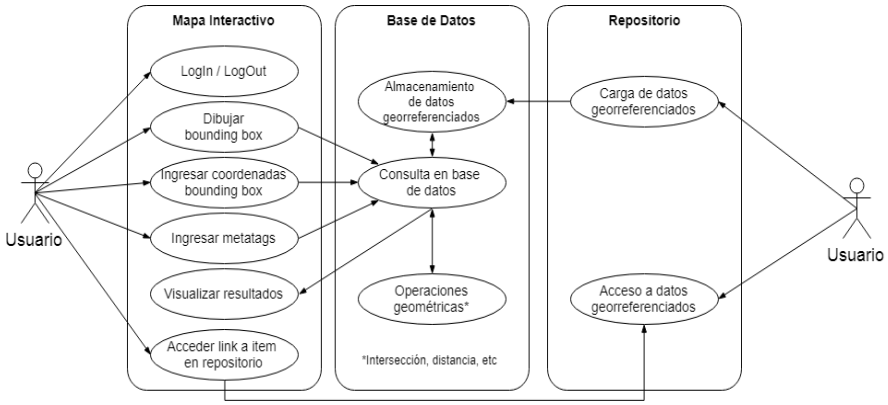

Figura 2. Caso de uso del sistema y relaciones entre componentes.

2.4 Repositorio

Se utiliza DSpace como repositorio de datos primarios. DSpace es un software de código abierto que provee herramientas para la administración de colecciones digitales. Actualmente es una de las aplicaciones más utilizada por universidades y organismos de investigación como RI [6]. DSpace organiza el repositorio en comunidades, colecciones e ítems. La comunidad es el nivel más alto de la jerarquía en el mismo, puede ser un instituto, un centro de investigación, un laboratorio, etc. Una comunidad también puede contener sub-comunidades y dentro de cada comunidad contiene una o más colecciones. A su vez una colección puede contener uno o más ítems, siendo un ítem la unidad mínima de información. Un ítem está compuesto por metadatos que describen su contenido y uno o más archivos que forman su contenido.

Para la descripción del contenido de un ítem se utilizan metadatos del estándar Dublin Core (“dc”) [7], algunos de ellos son obligatorios y los utiliza el mismo DSpace para su gestión. Por otro lado, permite incluir nuevos esquemas de metadatos para la descripción de los ítems almacenados y facilitar su búsqueda y recuperación.

Este software está escrito en Java y permite utilizar PosgtreSQL u Oracle como base de datos relacional para su gestión. Toda la información queda almacenada en la base de datos excepto los archivos que se registran en una serie de directorios administrados por DSpace.

2.4.1 Definición de Nuevos Esquemas de Metadatos

Además del esquema de metadatos “dc” que utiliza DSpace, se crea el esquema “geo” (Tabla 1) que permite registrar información del archivo, dentro de un ítem, que

contiene datos georreferenciados, la ubicación de sus campos, su formato, etc. para poder extraer los datos necesarios para su procesamiento y exportación.

Los dataloggers generan registros en distintos formatos según su fabricante y modelo, pero todos los dataloggers utilizados cuentan con la opción de bajar sus datos en formato de archivos csv[1].

Tabla 1. Esquema de metadatos geo

geo.FileName	Nombre de archivo que contiene datos de georreferencia
geo.ColumnLongitude	Nº de columna que contiene la longitud en grados decimales
geo.ColumnLatitude	Nº de columna que contiene la latitud en grados decimales
geo.ColumnDataTime	Nº de columna que contiene la fecha y hora
geo.FormatDataTime	Formato utilizado para extraer la fecha y hora
geo.Delimiter	Delimitador utilizado como separador de columnas
geo.Resolution	Resolución utilizada para representar la trayectoria

Dado el tipo de contenido de los ítems, de carácter biológico, también se creó el esquema de datos Darwin Core (“dwc”) [8] con todos los metadatos que provee la versión Darwin Core 1.4.

2.4.2 Configuración de las Colecciones que Registran Dataloggers

Para la carga de ítems con datos de dataloggers se crearon las colecciones donde se almacenan los mismos, se definió el subconjunto de metadatos de “dc” y se configuró la utilización de metadatos de los esquemas “geo” y “dwc”.

2.4.3 Tarea para el Procesamiento y Exportación de Datos

DSpace provee facilidades para lanzar distintos tipos de procesamientos sobre el contenido de un repositorio. Dentro de DSpace se lo conoce como “Sistema de Curación” y a cada proceso como “Tarea”. Las mismas son configurables y pueden operar sobre cualquier objeto de DSpace, por ejemplo el sitio entero, una comunidad, una colección, un ítem, etc.

Para realizar el procesamiento y exportación de datos de un ítem que contiene un archivo georreferenciado se desarrollaron los módulos (en Java) siguiendo los

[1] csv: (del inglés comma-separated values) archivo de texto que representa datos en forma de tablas en donde las columnas se separan por una coma (u otro carácter definido previamente)

estándares que sugieren los desarrolladores de DSpace [9], por otro lado se creó una nueva tarea que la vincula con los módulos desarrollados.

Los módulos mencionados seleccionan el archivo que contiene datos georreferenciados dentro de un ítem, ubica las columnas y formatos utilizados y extraen datos del mismo, a partir de los datos cargados en "geo".

Los datos son exportados a una tabla, llamada "derrota" que contiene toda la información acerca de los viajes de los especímenes, la misma se agregró en la base de datos PostgreSQL que utiliza DSpace para su gestión. En este caso se inserta un registro por cada ítem exportado, guardando la referencia del ID del ítem procesado.

2.4.4 Workflows de las Colecciones que Registran Dataloggers

Otra de las opciones que posee DSpace es la posibilidad de definir workflows para la carga de ítems, pudiendo definir usuarios/grupos que inician el proceso de carga, revisores, etc. hasta que un ítem queda definitivamente cargado y disponible dentro de la colección.

En este caso se usó esta opción para lanzar la tarea de procesamiento y exportación de datos ni bien el ítem queda cargado en la colección. En caso de errores (por ejemplo el formato especificado en "geo" de una columna es incorrecto) o de cambios en el ítem, se puede lanzar nuevamente la tarea para ese ítem.

DSpace también permite lanzar la tarea para todos los ítems dentro de una colección permitiendo generar nuevamente todos los registros en la tabla "derrota" para esa colección.

2.5 Base de Datos

PostgreSQL [10] es un sistema de gestión de bases de datos relacional orientado a objetos (ORDBMS) de código abierto que usa y amplía el lenguaje SQL, es transaccional y compatible con ACID.

PostGIS [11] es una extensión espacial de código abierto para PostgreSQL. Agrega soporte para objetos geográficos que permiten que las consultas de ubicación se ejecuten en SQL y sigue las especificaciones "Simple Features Specification For SQL" de Open Geospatial Consortium (OGC) [12].

En nuestra implementación se crea la tabla "derrota", en la base de datos de DSpace, con los campos indicados en la Tabla 2.

Tabla 2. Campos de la tabla derrota.

Nombre	Tipo	Descripción

id	bigint	Identificador autoincremental
name	text	Descripción del viaje
geom	geometry(Geometry,4326)	Objeto tipo ST_Geometry con las coordenadas que definen la ruta, utilizando SRID 4326
id_item	uuid	Foreign key a la tabla de items de DSpace
dist_mayor	character(10)	Distancia mayor a la colonia en km
tiempo	character(10)	Duración del recorrido en horas

Al realizar una consulta, se genera un polígono rectangular utilizando la función "ST_MakeEnvelope". Luego, mediante la función "ST_Intersects" se lleva a cabo la intersección entre el polígono rectangular recién generado y todas las derrotas existentes en la base de datos. Un ejemplo de consulta SQL podría ser:

```
SELECT distinct handle.handle,st_astext(derrota.geom),der-
rota.name,ST_Length(derrota.geom::geography)/1000,
derrota.dist_mayor,derrota.tiempo
FROM handle,derrota
WHERE derrota.id_item = handle.resource_id AND
ST_Intersects(derrota.geom,ST_MakeEnvelope(${bbox_left},
${bbox_bot},${bbox_right},${bbox_top},${srid}));
```

En la consulta SQL se relacionan las tablas derrota y handle. La tabla handle almacena datos de los ítems del repositorio DSpace.

Cada registro encontrado está compuesto de:

- **handle.handle:** Identificador del ítem en DSpace. Con este dato se genera el enlace para visualizar el ítem en DSpace.
- **st_astext(derrota.geom):** Conversión de objeto de geometría (LINESTRING) a secuencia de coordenadas, para su representación gráfica.
- **derrota.name:** Descripción del viaje.
- **ST_Length(derrota.geom::geography)/1000:** Conversión de objeto de geometría (LINESTRING) a distancia, expresada en km.
- **derrota.dist_mayor:** Distancia mayor a la colonia, expresada en km.
- **derrota.tiempo:** Duración del recorrido, expresada en horas:minutos:segundos.

Ambas tablas se relacionan por medio de las foreign keys "derrota.id_item" y "handle.resource_id" (primary key de dspaceobject.uuid) y por medio de la intersección del objeto de geometría que define el recorrido del animal con el polígono rectangular creado con los vértices del rectángulo definido en la interfaz gráfica.

2.6 Mapa Interactivo

La aplicación de visualización y consulta se desarrolló utilizando la biblioteca de JavaScript Leaflet. La misma es ampliamente utilizada en aplicaciones Web que interactúan con mapas, cuenta con una gran variedad de funcionalidades, buen rendimiento, está bien documentada y es fácil integrarla con otras bibliotecas que extienden su funcionalidad [13].

Para el desarrollo se aplicó la metodología incremental, comenzando con un mapa interactivo básico y agregando nuevos elementos que amplían su funcionalidad.

Se comenzó con la codificación que permite la visualización de un mapa base. Luego se agregó un elemento de control que permite trazar rectángulos el cual se utiliza para realizar búsquedas por coordenadas y otro elemento para borrarlos. Ambos elementos se incluyen en un plugin adicional llamado Leaflet.Draw.

También se extendió la funcionalidad agregando búsqueda y filtrado de resultados por metadatos, por medio de elementos de control, que permite el ingreso de palabras clave.

La funcionalidad de búsqueda se resuelve con una solicitud AJAX, que envía las latitudes y longitudes de dos vértices opuestos que definen el rectángulo, se arma una consulta SQL que se ejecuta sobre la base de datos del repositorio y recibe como respuesta una cadena de datos en formato GeoJSON [14]. Como resultado se obtiene un conjunto de elementos geométricos, los cuales son interpretados por Leaflet y representados en el mapa.

Finalmente se agregan elementos de código que proveen una mejor visualización de datos y funcionalidades secundarias:

- **Rectángulo manual:** permite definir el tamaño y posición exacta de un rectángulo mediante el ingreso de las coordenadas del mismo.
- **Coloración de resultados:** para mejorar la identificación, se asignan colores aleatorios a cada resultado a mapear.
- **Ventana emergente de un ítem:** a cada elemento dibujado en el mapa, se le agrega una ventana emergente que contiene datos básicos del ítem en el repositorio que representa, así como un link al mismo.
- **Listado de resultados:** se muestra un menú emergente en un lateral con el listado de los resultados obtenidos, para una rápida visualización e identificación de los mismos que además permite seleccionar y acceder a los datos en el repositorio.
- **Control de acceso:** para proteger y restringir el acceso a los datos, se agrega una pantalla de Login, la cual provee control y limitación de las colecciones a ser incluidas en las búsquedas, según los privilegios asignados a cada usuario.

En la figura 3 se muestra los resultados de una consulta utilizando la interfaz desarrollada.

Figura 3. Ejemplo de búsqueda.

3 Resultado y Trabajos Futuros

Se ha desarrollado un sistema que permite extender las funcionalidades de DSpace, donde es posible almacenar datos primarios georreferenciados con formatos heterogéneos.

La solución propuesta da respuesta a una necesidad concreta de gestión de datos primarios a un grupo de investigación. Cabe destacar que la misma es válida para todo grupo de investigación que trabaje con grandes volúmenes de datos generados por dataloggers, ya que requerirá de una organización y manipulación de datos y metadatos similar.

Este sistema se puso en producción en agosto de 2016 y se presentó a todos los Institutos del CCT CONICET-CENPAT quienes mostraron interés en aplicarlo en proyectos que requieren la gestión de datos primarios.

Los trabajos futuros en este proyecto incluyen la extensión de su funcionalidad para otras áreas temáticas de la ciencia, la generación de reportes y exportación de datos y mejoras de la interfaz de visualización y consulta.

Se pretende incorporar tecnologías de Web Semántica a las búsquedas en el marco del proyecto de investigación del LINVI “Infraestructura de Acceso a Datos Primarios con aporte de semántica en Repositorios Digitales”.

Referencias

1. Chapman AD, Wieczorek J. Guide to best practices for georeferencing. Copenhagen: Global Biodiversity Information Facility. 2006; 1–77.
2. Boehme L, Kovacs K, Lydersen C, Nøst OA, Biuw M, Charrassin JB, et al. Biologging in the global ocean observing system. Proceedings of OceanObs 09: Sustained Ocean Observations and Information for Society (Vol 2), Venice, Italy, 21-25 September 2009, Hall, J , Harrison DE & Stammer, D , Eds , ESA Publication WPP-306. 2010; Available: http://epic.awi.de/21347/1/Boe2009h.pdf

3. Block BA, Jonsen ID, Jorgensen SJ, Winship AJ, Shaffer SA, Bograd SJ, et al. Tracking apex marine predator movements in a dynamic ocean. Nature. 2011;475: 86–90.
4. Egenhofer MJ. Toward the semantic geospatial web. Proceedings of the 10th ACM international symposium on Advances in geographic information systems. ACM; 2002. pp. 1–4.
5. DSpace - A Turnkey Institutional Repository Application. In: Duraspace.org [Internet]. Available: https://duraspace.org/dspace
6. DuraSpace Registry - Duraspace.org. In: Duraspace.org [Internet]. Available: http://registry.duraspace.org/registry/dspace
7. Initiative DCM, Others. DCMI home [Internet]. 2013. Available: http://dublincore.org
8. Darwin Core - Darwin Core [Internet]. Available: http://rs.tdwg.org/dwc
9. Developer Guidelines and Tools - DSpace - DuraSpace Wiki [Internet]. Available: https://wiki.duraspace.org/display/DSPACE/Developer+Guidelines+and+Tools
10. Group T, Others. PostgreSQL: The world's most advanced open source database [Internet]. 2011. Available: https://www.postgresql.org
11. Developers P. PostGIS — Spatial and Geographic Objects for PostgreSQL [Internet]. Available: http://postgis.net
12. Simple Features SWG | OGC [Internet]. Available: http://www.opengeospatial.org/projects/groups/sfswg
13. WebAgafonkin V. Leaflet-an Open-source JavaScript Library for Interactive Maps. Leaflet Dev Blog Atom Np, 2015 Web. 2016; Available: https://leafletjs.com
14. Butler H, Daly M, Doyle A, Gillies S, Hagen S, Schaub T. The geojson format [Internet]. 2016. Available: https://tools.ietf.org/html/rfc7946